Propuesta de Tesis

Alumno: Damián Eliel Aleman
Directores: Juan Manuel Pérez y Santiago Kalinowski
Codirector: Agustín Gravano

16 de mayo de 2017

Introducción

La *lingüística de corpus* es un área que ha tenido desarrollo sostenido desde los años cincuenta, acompañando los avances que fue haciendo la tecnología. El problema principal con el que debe lidiar la disciplina es que resulta trabajoso generar un dataset que sea válido científicamente sin dedicar una enorme cantidad de recursos a su recolección.

En esta tesis queremos poder analizar sobre una base estadística ciertos fenómenos lingüísticos del habla hispana utlizando un dataset propio con textos recolectados de *Twitter*.

Este trabajo se realizará en colaboración con el Departamento de Investigaciones de la Academia Argentina de Letras, cuyo director, el Dr. Santiago Kalinowski, es uno de los directores de esta tesis.

Trabajo Previo

Actualmente, las herramientas para la detección de palabras con contraste léxico en distintas regiones consisten en cuestionarios como el de *Almeida* [1]. Estos cuestionarios están integrados por grupos temáticos centrales como la casa, la familia, la enseñanza, el cuerpo humano, etc. Sobre cada grupo temático se les indicaba a las personas entrevistadas un repertorio de palabras por cada noción, para ver si las conocían y con que frecuencia se las usaba. Con este trabajo planteamos cambiar el paradigma y detectar automáticamente las palabras usadas en distintas regiones y sus frecuencias.

Uno de los corpus lingüísticos del español más reconocidos es el *corpes XXI*[2], creado por la Real Academia Española con una distribución de 25 millones de formas por cada uno de los años comprendidos en el periodo 2001 a 2012. Sin embargo, dicho corpus tiene dos desventajas importantes: por un lado, la cantidad de palabras de América Latina están subrepresentados ya que el $65,70\,\%$ de las palabras del dataset provienen de textos de España y $34,30\,\%$ de los demás países hispanoparlantes . Por otro lado, uno no

dispone de todo el dataset, sino que solamente se pueden hacer las consultas de su página web. Estas consultas están limitadas en cuanto a la cantidad de solicitudes y a las funcionalidades que estas proveen.

Objetivo de la tesis

El objetivo del presente trabajo sería, en primer lugar, construir un dataset[3] sobre textos en español de Argentina extraídos de *Twitter*[4]. Sobre este dataset, además del texto nos interesará tener información de quién lo escribe y en particular de qué parte de la Argentina proviene.

Teniendo el conjunto de datos, se realizará un análisis estadístico sobre el listado de frecuencias de palabras, para poder detectar neologismos y contrastes de uso en distintas regiones ya sea a nivel provincia, o entre conjuntos de provincias. Luego, se analizará la correlación entre las frecuencias de palabras de las regiones dialectales provistas por *Vidal de Batini*[5].

Un éxito en esta etapa, aunque sea parcial, abre infinidad de posibilidades de trabajo con la lengua en Internet que tiene muchas aplicaciones más allá del interés puramente lingüístico: sondeos de opinión, publicidad, detección de tendencias en distintas áreas (deporte, cine, televisión, consumo de bienes, etc.).

Método

Para extraer los tweets se utilizará la librería de *python* llamada *tweepy*. A través de la librería se hará una búsqueda de usuarios de forma localizada para después extraer tweets de estos. Luego se procesará el texto, para limpiar aquellas secuencias que no nos interesan para el análisis lingüístico, como los números, emoticones, o signos de puntuación entre otros. Finalmente haremos un listado de frecuencias de palabras[6] por provincia, y por las regiones dialectales obtenidas del trabajo de *Vidal de Batini*. Sobre estos listados realizaremos análisis estadísticos para detectar las palabras con contrastes significativos.

En esta tesis se plantearán varios desafíos computacionales para la creación del corpus y para el posterior análisis estadístico de las distribuciones de palabras. Como resultado, realizará mejoras significativas a la metodología empleada por la Academia Argentina de Letras al estudio del uso del español en nuestro país.

Referencias

[1] Almeida Manuel, Vidal Carmelo, *Variación socioestilística del léxico, Boletín de Filología ,35.1 Pág. 50-64* 1995

[2] Real Academia Española, *Banco de datos (CORPES XXI) [en línea]. Corpus del Español del Siglo XXI (CORPES). <http://www.rae.es>*

[3] McEnery, Tony, Andrew Hardie. *Corpus linguistics: method, theory and practice* 2012

[4] Kilgarriff, Adam, Grefenstette Gregory, *Introduction to the special issue on the web as corpus*, Computational linguistics 29.3 (2003): 333-347.

[5] Vidal de Battini, Berta Elena, *El español en la Argentina* 1964

[6] Baayen, R. Harald *Word frequency distributions. Vol. 18. Springer Science & Business Media* 2001